Poster I-24

Computing the Most Conserved Protein

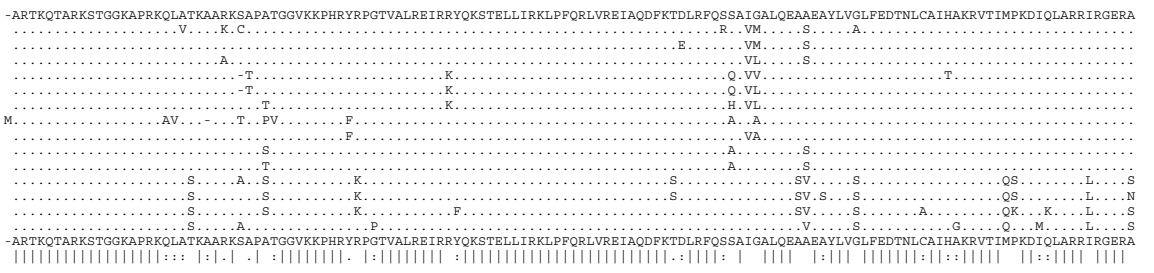
Goldberg, David, Bern, Marshall*

Scripps PARC Institute for Advanced Biomedical Sciences, Palo Alto, CA, USA

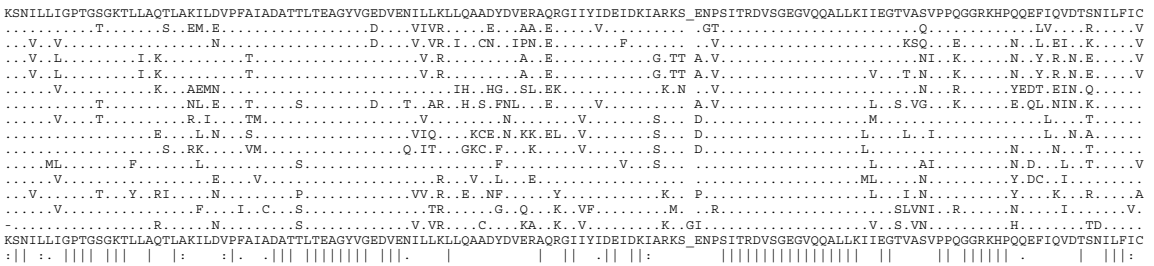
We introduce an innovative computational approach to data mining the ever growing collection of complete genomes. We ask the question "what is the most conserved protein?", introduce an algorithm for computing the answer that is faster, more flexible and more meaningful than can be done with BLAST, and discuss several applications for the results.

To get meaningful results, all proteins must be compared on an apples-to-apples basis. We do this by fixing a window size *k* (typically 50-500 amino acid residues) and comparing all proteins using only a *k*-mer from each protein. This would be very difficult to do efficiently with BLAST. Our algorithm takes a set of organisms and finds the most conserved (orthologous families of) proteins, ranking them from most to least conserved. The conservation measure for each family is the "diameter" of the cluster formed by selecting a *k*-mer from each protein in the family. The algorithm computes the *k*-mers that give the maximum amount of conservation.

We have run the algorithm over the complete bacterial genomes in Genbank. The following figure gives a glimpse of our results. For comparison it shows Histone H3 from a wide range of eukaryotes. H3 is a 135-residue protein, and so we show the most conserved 135-mer over a wide range of bacteria for ClpX, one of the most highly conserved proteins in that window size.



Histone H3 for various eukaryotes (including single-celled protists)



ATP-binding subunit ClpX for various eubacteria. This segment starts about 100 residues from the amino end.

Some applications of an algorithm for computing protein conservation are:

- *Phylogeny*. Phylogenetic trees are traditionally built with rRNA. Using highly conserved proteins instead yields higher quality trees (*submitted for publication*).
- *Finding characteristic proteins for a clade*. The algorithm can be run over specific clades, and the proteins which are significantly more highly conserved compared to a superclade can be identified.